Non-Rigid Objects Recognition: Automatic Human Action Recognition in Video Sequences

Mehrez Abdellaoui[1], Ali Douik[1] and Kamel Besbes[2]
[1]National Engineering School of Monastir,
[2]Faculty of Sciences of Monastir,
Tunisia

1. Introduction

Non-rigid objects recognition is an important problem in video analysis and understanding. It is nevertheless a challenging task to achieve due to the properties carried out by the non-rigid objects, and is more complicated by camera motion as well as background variation. Human body recognition in video sequences is the best application of the non-rigid objects recognition due to the large capacities of the human body in doing actions and poses. These difficulties prohibit practical attempts toward conceiving a robust global model for each action class. Human body recognition is highly interesting for a variety of applications: detecting relevant activities in surveillance video, summarizing and indexing video sequences. It relies, however, on the interpretation of the body movements and classifies them in different events.

A considerable amount of previous work has addressed the question of human action categorization and motion analysis. One line of work is based on the computation of correlation between volumes of video data (Efros et al., 2003). Another popular approach is to track body parts at first and then uses the obtained motion trajectories to perform action recognition (Ramanan & Forsyth, 2004). The robustness of the approach is highly dependent on the tracking system. Alternatively, researchers have considered the analysis of human actions by looking at video sequences as space-time intensity volumes (Bobick & Davis, 2001). Some researchers have also explored unsupervised methods for motion analysis such as hierarchical dynamic Bayesian network model (Hoey, 2001; Zhong et al., 2004). Another approach uses a video representation based on spatiotemporal interest points (STIPs). In spite of the existence of a fairly large variety of methods to extract interest points (IPs) from static images Harris corner detector (Harris & Stephens, 1988), Scale invariant feature transform (Lowe, 1999), Salient regions (Kadir & Brady, 2003) …, less work has been done on STIPs detection in videos. In 2005, Laptev (Laptev, 2005) present a STIPs detector based on the idea of the Harris IPs operators. They detect local structures in space-time where the image values have significant local variations in space and time dimension. IPs extracted with such methods had been used as features for human action classification. These points are particularly interesting because they focus the initial information contained in any image in a few specific points. The integration of the time component can perform filtering on the IP and keep only those who also have a temporal discontinuity.

We propose in this chapter a motion analysis and classification approach to learn and recognize human actions in video, taking advantage of the robustness of STIPs and the unsupervised learning approaches. Experimental results are validated on KTH human action database (Schuldt et al., 2004), and ATSI Human Action Database (see Figure 1). Results are compared to recent works on the human motion analysis and recognition.

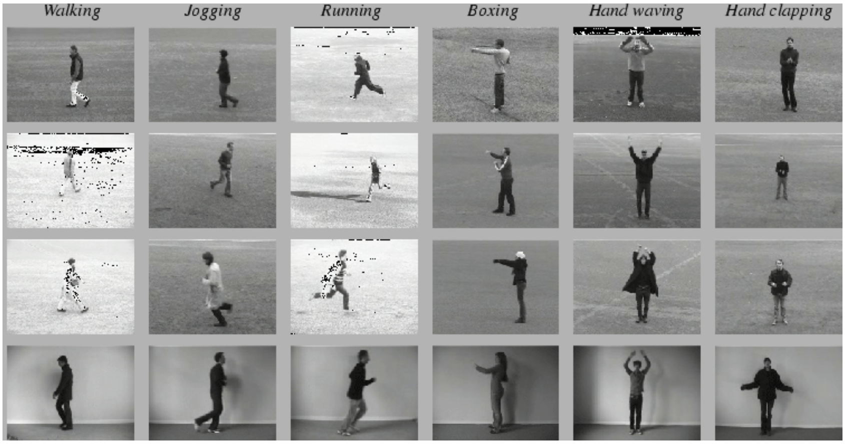


Fig. 1. Samples from the KTH human action database

2. Spatio-temporal interest points

2.1 Presentation

Interest Points in a bitmap image are defined as pixels with maximum variations of the intensity in the local neighbourhood. These pixels represent corners, intersections, isolated points and specific points on image texture. This definition can describe the Spatio-temporal Interest Points (STIPs) when considering a video sequence instead of the image. Consequently, we deduce that STIPs can be defined as pixels with significant changes in space and time. It can represent irregular movements of the human body such as bending elbows or knees, moving limbs. Whereas, uniform movement such as moving a hard object does not generate any STIP. Video sequences are represented as a 3D function over two spatial dimensions (x, y) and one temporal dimension t. Many detectors can be used such as: Laptev et al. detector (Laptev & Lindeberg, 2004); Dollàr et al. detector (Dollár et al., 2005); FAST-3D detector (Koelstra et al., 2009); and Oikonomopoulos et al. detector (Oikonomopoulos et al., 2006).

2.2 Laptev et al. detector

The Laptev et al. theory (Laptev & Lindeberg, 2004) is based on the Harris operator (Harris & Stephens, 1988) that had shown good performances interest points detection in static images. The operator extension over the spatiotemporal domain makes the spatio-temporal interest points detection possible. This extension consists of a search of points that maximize the local variation of image values simultaneously over the spatial dimensions and the temporal dimension. According to Laptev et al., a video sequence can be represented as a function $f : R^2 \times R \to R$ over two spatial dimensions (x, y) and one temporal dimension t. The Local space time features are defined as 3D blocks of the sequence containing variations in space and time.

The scale-space representation $L : R^2 \times R \times R_+^2 \mapsto R$ is generated by the convolution of f with a separable Gaussian kernel g (p ; Σ) (1). Where p is spatiotemporal position vector $p = (x, y, t)^T$, the parameters σ^2 and τ^2 of the covariance matrix correspond to the spatial and temporal scale parameters respectively and define spatiotemporal extension of the neighbourhoods.

$$g(p;\Sigma) = \frac{1}{\sqrt{(2\pi)^3 \det(\Sigma)}} e^{-\frac{\left(p^T \Sigma^{-1} p\right)}{2}} \quad \text{and} \quad \Sigma = \begin{pmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \tau^2 \end{pmatrix} \tag{1}$$

A spatiotemporal second-moment matrix (2) is defined in terms of spatiotemporal gradients and weighted with a Gaussian window function.

$$\begin{aligned} \mu(\cdot;\Sigma) &= g(\cdot;\Sigma) * \left(\nabla L(\cdot;\Sigma) \left(\nabla L(\cdot;\Sigma) \right)^T \right) \\ &= g(\cdot;\Sigma) * \begin{pmatrix} L_x^2 & L_x L_y & L_x L_t \\ L_x L_y & L_y^2 & L_y L_t \\ L_x L_t & L_y L_t & L_t^2 \end{pmatrix} \end{aligned} \tag{2}$$

The spatiotemporal second-moment matrix μ, considered also as a structure tensor, is interpreted in terms of eigen values. This fact makes the distinguishing of image structures possible with variations over one, two and three dimensions. Three-dimensional variation of f corresponds to image points with non-constant motion. Such points can be detected by maximizing the three eigen values λ1, λ2, λ3 of μ over space and time.

STIP detection is realized by the extension of the Harris operator H into the spatiotemporal domain (3). Detection is based on points with high eigen values.

$$H = \det(\mu) - k \cdot \mathrm{trace}^3(\mu) = \lambda_1 \cdot \lambda_2 \cdot \lambda_3 - k \cdot \left(\lambda_1 + \lambda_2 + \lambda_3 \right)^3 \tag{3}$$

Local maxima of H correspond to points with high values λ1, λ2, λ3 (λ1 ≤ λ2 ≤ λ3). H can be written as equation (4), where α = λ2/λ1 and β = λ3/λ1.

$$H = \lambda_1^3 \left(\alpha\beta - k \left(1 + \alpha + \beta \right)^3 \right) \tag{4}$$

From the requirement H ≥ 0, we get the condition represented by (5).

$$k \le \alpha\beta \big/ \left(1 + \alpha + \beta \right)^3 \tag{5}$$

And it follows that for perfectly isotropic image structures ($\alpha = \beta = 1$), k assumes its maximum possible value kmax = 1/27. For sufficiently large values of k ≤ kmax, positive local maxima of H will correspond to space-time points with similar eigen values $\lambda 1$, $\lambda 2$, $\lambda 3$. Consequently, such points indicate locations of image structures with high spatiotemporal variation and can be considered as positions of local spatiotemporal features. As k in (3) only controls the local shape of image structures and not their amplitude, the method for local features detection is invariant with respect to the affine variation of image brightness.

2.3 Dollàr et al. detector

Compared to Laptev detector, Dollàr et al. detector (Dollàr et al., 2005) it produces dense features that can significantly improve the recognition performance in most cases. It uses two separate filters in spatial and temporal directions: 2-D Gaussian filter in space components and 1-D Gabor filter in time component.

A response function of the form (6) is obtained, where g is the 2D Gaussian kernel applied along the spatial dimensions of the video and h_{ev} (7) and h_{od} (8) are a quadrature pair of 1D Gabor filters applied in the temporal dimension.

$$R = (I * g * h_{ev})^2 + (I * g * h_{od})^2 \tag{6}$$

$$h_{ev}(t; \tau, \omega) = -\cos(2\pi t\omega)\ e^{-t^2/\tau^2} \tag{7}$$

$$h_{od}(t; \tau, \omega) = -\sin(2\pi t\omega)\ e^{-t^2/\tau^2} \tag{8}$$

The detector responds best to complex motions made by regions that are distinguishable spatially, including spatio-temporal corners, but not to pure translational motion or motions involving areas that are not distinct in space. Local maxima of the response function R are selected as interest points, and cuboids are extracted, which are the windowed pixel values around the interest point in the spatial and temporal dimensions.

2.4 The FAST-3D detector

The FAST-3D spatio-temporal detector, developed by (Koelstra et al., 2009), is inspired from the FAST detector (Features from Accelerated Segment Test detector). Instead of using a circle around each pixel (x, y, t), Koelstra et al considered the set C of the 26 directly neighbouring pixels to (x, y, t) in a 3D space-time neighbourhood. STIPs detection is correctly done even when videos are transformed by zoom, rotation or MPEG compression.

2.5 Laptev detector Implementation

The algorithm was applied to sequences of different types of video sequences for detecting the STIP. The application of the algorithm is made through two executable files "stipdet.exe" and "stipshow.exe". The first file corresponds to the detection algorithm STIP and the second for showing the detected STIPs on the sequences.

The implementation of the first program generates a text file with space-time coordinates of the tracks (x, y, t). The second program displays STIPs detected on the images of the video sequence. Video sequences are processed using Matlab with a single variable representation. The three-dimensional tensors represent properly video sequences. Figure 2 shows the detected STIPs in different video frames' samples from the KTH human action database. The three components are x (height) y (widths) and t (time axis). This representation makes possible the STIPs neighborhood search in space-time domain.

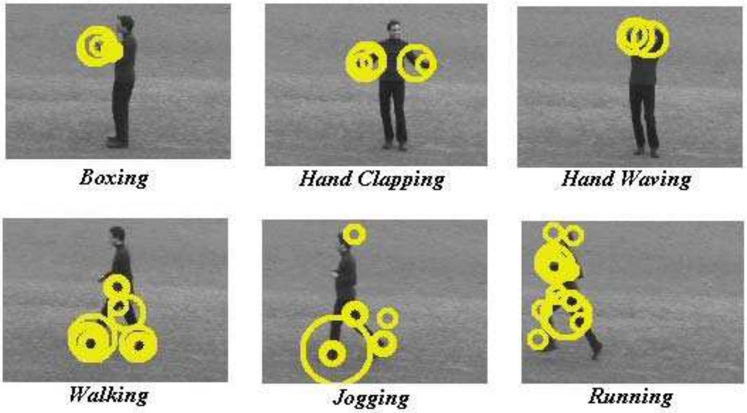


Fig. 2. Detected STIPs in different samples from the KTH human action database

Among all detected STIPs from video sequences, there are usually motion noises from the non uniform background that do not contribute to the action motion. In fact, those points normally make the modeling computation much harder and in some cases might completely distract the core parts of the action. In order to filter out these irrelevant elements, we consider only STIPs that coincide with the dilated shape of the human body.

The tensor elements contain gray level values of pixels in each frame of the video sequence. The criteria developed by Laptev et al. are applied on tensors and STIPs detected are pixels with maximum values in local neighborhoods and this by maximizing the criterion H. Figure 3 shows the structure of the tensor with the three axes.

The STIP detected by the Laptev algorithm have interesting properties including their stability to geometric transformations. Other robustness properties of the STIPs can be determined. These properties are related to noise from video sequences, such as impulse noise, contrast changes, quick movement of the camera and the MPEG compression effects. Several studies have been done in this area. Lejeune-Simac et al. (Lejeune-Simac et al., 2010) present a comprehensive study of the robustness of the detector STIPs various effects of noise from video sequences.

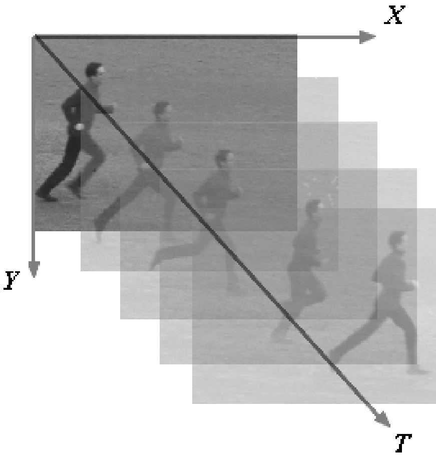


Fig. 3. Reference axes (x,y,t) representation on a video sequence from KTH human action database

3. Motion analysis approach

To analyse motion we defined different parameters based on STIP detection using Laptev detector. The first one calculates the number of STIP in the sequence, whereas the second is the “activity” function that evaluates the evolution of STIP during the sequence and the third parameter analyses the position of STIP points comparing to reference one associated to the body in movement.

3.1 Number of STIPs in a sequence

Human body movements can be differentiated by a quantitative survey on STIPs detected. Thus, an algorithm was developed with a purpose to calculate STIPs number for each sequence from different human body motion databases. This algorithm leads to interesting results. Indeed, STIPs number is high for fast movements like (running, jogging, jumping). Other movements made only by the arms (boxing, hand clapping or hand waving) lead to low STIPs number. Table 1 shows the evolution of the STIPs average number in 100 frames sequences (4 seconds of video) for each movement class. The algorithm was tested on 450 sequences from KTH database (75 for each movement).

Movement	STIPs average number per 100 frames
Running	685
Jogging	463,33
Walking	313,33
Hand waving	145
Hand clapping	114
Boxing	82

Table 1. Number of STIPs evolution for KTH human action database.

These statistics show that STIPs number depends directly of the movement realized. Indeed, running and jumping movements have high STIPs number however boxing and hand waving have a low STIPs number. Therefore we conclude that STIPs number in a sequence is an important parameter in human movements' recognition. To emphasize this study we present in the following section the evolution of STIPs in time by the "Activity" function.

3.2 Activity function

Evolution of the STIPs number in a sequence is an important factor in human motion recognition. To synthesize this criterion we have used the "Activity" function. This function was defined by Laganière et al. (Laganière et al., 2008) as the number of pixels that are modified between two consecutive frames in a video sequence. Hence, frames that correspond to local maxima of the "Activity" function are the scenes of major movements. We have changed the "Activity" to fit our research, so we defined it as STIPs number in each frame of the sequence. The evolution of this number can lead us to recognize the type of movement made by detecting its local maxima which are the locations of large amounts of movement and its distribution that indicates the positions of these quantities in time scale. In Figure 4, we present the activity function applied to samples of sequences from KTH database.

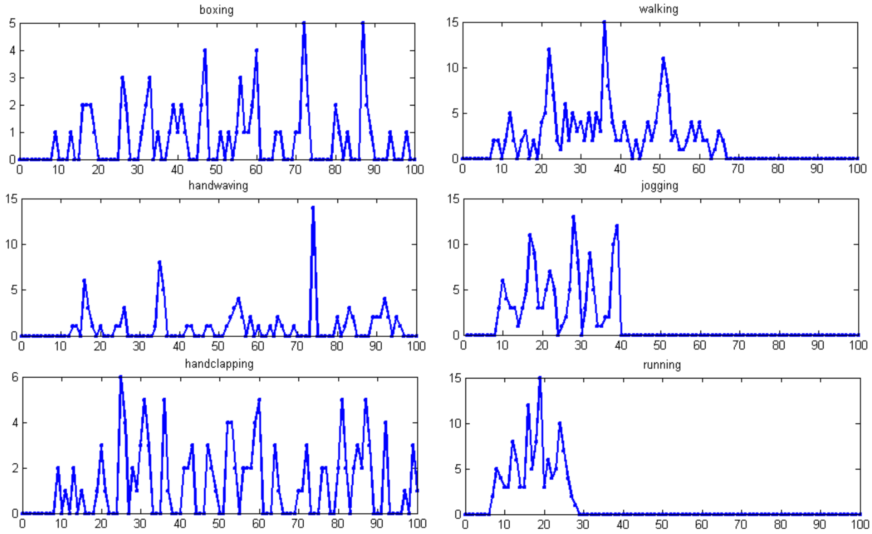


Fig. 4. Application of the Activity function on samples from KTH human action database.

The curves in Figure 4 have repetitive peaks. These peaks are local maxima of the activity function and can be regarded as major movement's events in each class. From this analysis we can extract important information about the class of the movement performed. The curves obtained are so noised. This is caused by non significant STIPs detected and which appear between local maxima. To resolve this problem we applied a smoothing algorithm on curves to accentuate the peaks and eliminate the STIPs values between the local maxima. The smoothing was done on segments of frames by adding the STIPs detected in an interval [n-2, n+2] where n is the time of the local maxima of the STIPs. Figure 5 shows the application of smoothing algorithm on the activity function curves for samples from the KTH human action database.

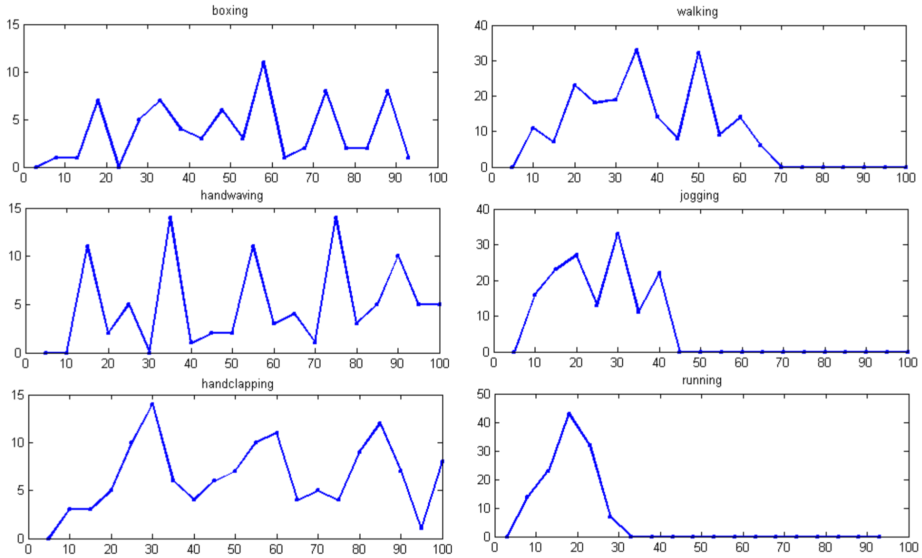


Fig. 5. Application of the smoothing algorithm on the activity function curves for samples from the KTH human action database.

We note that smoothing reduces the activity function noise and increases the local maxima values of the curves. To detect the locations of local maxima, a Gaussian model is fitted to the activity function. This model leads to the determination of the number of the local maxima and their time in a sequence. In addition, it contributes for motion recognition when considering the parameters of the used Gaussian model in the classification algorithm.

The value of the global maximum is deduced to detect movement with only one global maximum. Figure 6 shows the application of the Gaussian model to activity function on sequences taken from the KTH human action database (from left to right and row-wise of the Figure we have the actions of, boxing, walking, hands waving, jogging, hands clapping and running).

In Table 2, the number of local maxima is shown, their mean value and the global maximum value for different action classes taken from the KTH human action database. We note that the number of local maxima is the number of repetitions in a human movement such as walking or hand clapping. For fast movements such as running the smoothing algorithm reduces the number local maxima to one and extracts a single global maximum. The local maxima average value is a significant parameter in the classification of human movements. We note that the movements made only by arms such as: Boxing, Hand waving and Hand clapping have values lower than those achieved by the whole body such as: Running, Jogging and Walking. The global maximum can contribute to the classification since its values are different from one to another class of motion.

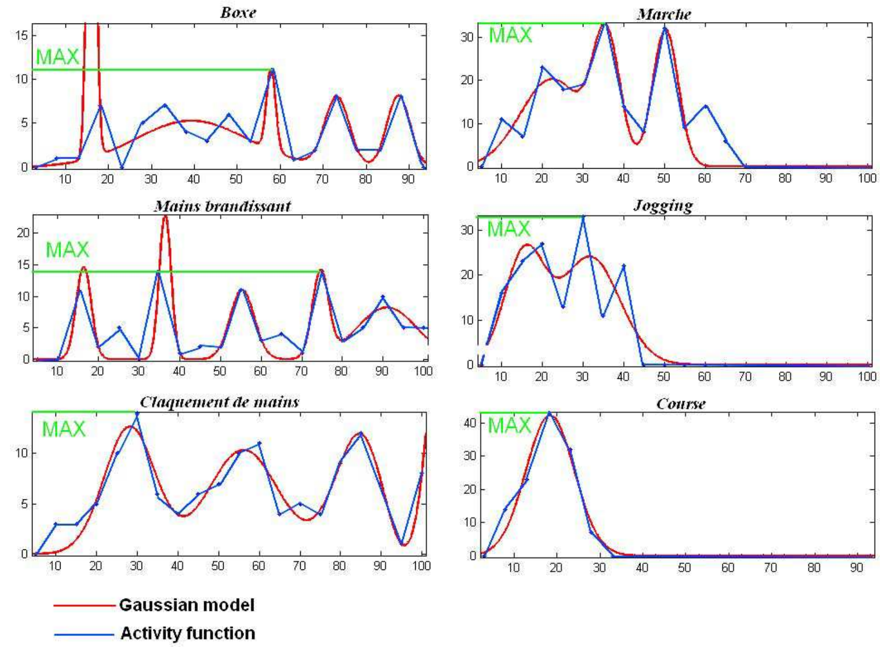


Fig. 6. Application of the Gaussian model to activity function on sequences taken from the KTH human action database

Action	Local maxima number	Local maxima mean value	Global maximum value
Running	1	42	43
Jogging	2	29	33
Walking	3	28	33
Hand waving	5	12	14
Hand clapping	3	12,33	14
Boxing	5	7,4	11

Table 2. Number of local maxima, mean value and the global maximum value for different action classes taken from the KTH human action database

The use of the activity function allows the tracking of the STIPs number in time. Its evolution has been modeled by a Gaussian model to extract its local maxima. This study can contribute to human motion recognition. Another important feature can be used. It consists on the spatiotemporal boxes associated to human body parts.

3.3 Spatiotemporal boxes

STIPs are the most significant motion locations in video sequences. Most of the STIPs are located at the most valuable human body parts such as knees, elbow joints, the moving limbs. Boxes containing STIPs called as “Spatiotemporal Boxes” can be considered as important information to describe the actions and to differentiate between them. Spatiotemporal boxes containing detected STIPs are the most shining regions to describe human motion. The boxes size can be effective information to differentiate between motion done only by hands and the full body motion (see Figure 7).

For all STIPs belonging to the same image, we determine their spatial coordinates (x1, y1) (x2, y2), ..., (xn, yn) in the image reference. The spatiotemporal boxes can be described by a rectangle between points (x_{Left}, y_{Top}) and (X_{Right}, y_{Bottom}) these coordinates are determined by reference to the following equations.

$$\begin{aligned} x_{Left} &= \min(x_1, x_2, \ldots, x_n) - r \\ y_{Top} &= \min(y_1, y_2, \ldots, y_n) - r \\ x_{Right} &= \max(x_1, x_2, \ldots, x_n) + r \\ x_{Bottom} &= \max(y_1, y_2, \ldots, y_n) + r \end{aligned} \tag{9}$$

r is the extension radius of the spatiotemporal boxes. Figure 7 shows spatiotemporal boxes detected on images taken from the KTH human action database.

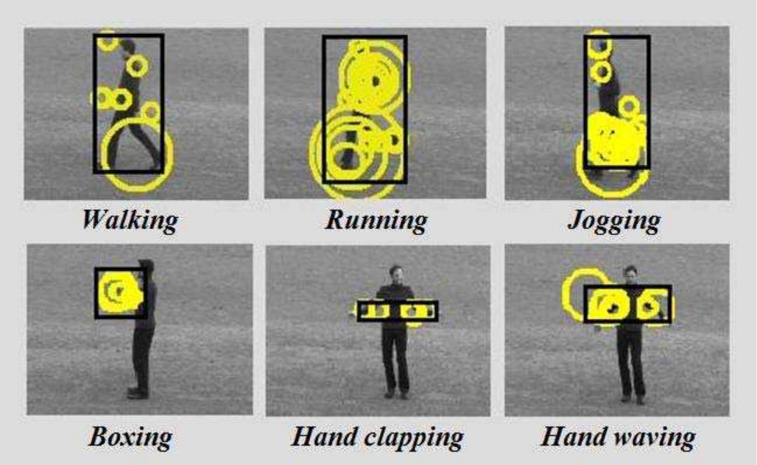


Fig. 7. Spatiotemporal boxes detected on images taken from the KTH human action database

Considering motion done using full body, we classified STIPs points in two parts, High body part STIPs (H-STIP) and Low body part STIPs (L-STIP). To achieve this classification we detected the centroid of the body silhouette in all frames of the sequence. Points located above centroid are classified in H-STIP and points below centroid are classified in L-STIP as shown in Figure 8.

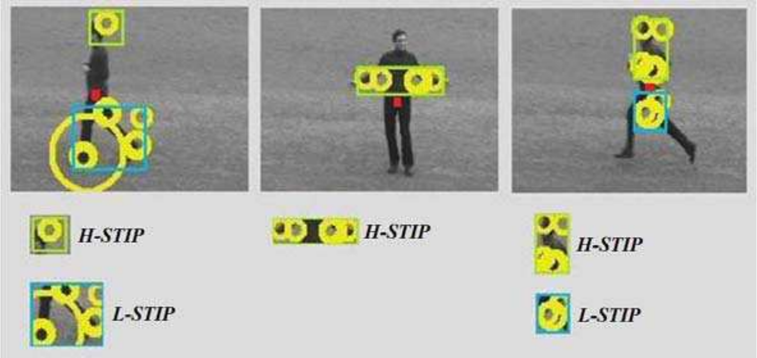


Fig. 8. Classification of H-STIP and L-STIP for action samples from KTH human action database

The evolution of H-STIP and L-STIP in time (see Figure 9) compared to centroid can be discriminative information to classify actions. In fact, actions containing H-STIP and L-STIP are Running, Jogging and Walking. On the other side, Boxing, Hand waving and Hand clapping contain only points of H-STIP type.

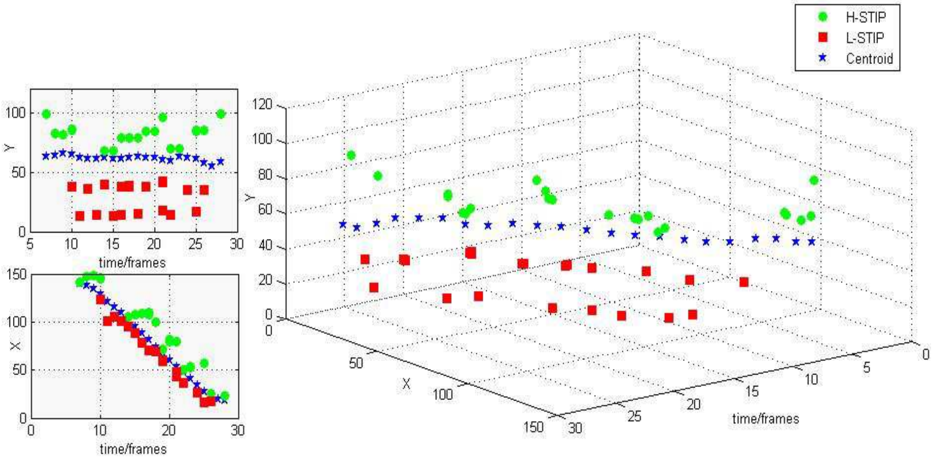


Fig. 9. An illustration of the evolution of H-STIP, L-STIP and Centroid in time for running action

4. Motion classification

To obtain fair judgement of the performances of the proposed approach, we compare our results with other human action recognition approaches using the same database. The performance of any approach is evaluated by measuring the accuracy of motion classification using a specified algorithm. Many algorithms can be used. The more used in will be described in the following subsections.

4.1 Probabilistic latent semantic analysis (pLSA)

It is a popular unsupervised method for learning object categories from interest point features and it was implemented based on Niebles et al. (Niebles et al., 2008). Histogram features of training or testing samples are concatenated to form a co-occurrence matrix which is an input of the pLSA algorithm.

4.2 Support Vector Machines (SVM)

Support Vector Machines (SVM) is one of the most popular classifier which has recently gained popularity within visual pattern recognition. In spatial recognition, local features have recently been combined with SVM in a robust classification approach. In a similar manner, Schuldt et al. (Schuldt et al., 2004) explored the combination of local space-time features and SVM and apply the resulting approach to the recognition of human actions.

4.3 Proposed algorithm

The classification algorithm is based on an unsupervised clustering algorithm K-MEANS The choice of this method is justified by the low running time and a priori knowledge of the number of classes K. The algorithm is based on a parameter vector V based on the criteria mentioned in previous sections. Table 3 shows the ranking of the parameters belonging to the vector V from the most to least significant paarameter.

Parameter	Feature
P1	Spatiotemporal box area
P2	Spatiotemporal box area/ Body bounding box area
P3	H-STIP existence (1 or 0)
P4	L-STIP existence (1 or 0)
P5	Distance between the spatiotemporal box centroid and the bounding box centroid
P6	STIPs Number /100 frames
P7	Global maximum value
P8	Local maxima number
P9	Mean value of local maxima
P10	Average value of the activity function variance
P11	Slope of the curve x=f(t) of the centroid

Table 3. List of parameters belonging to the vector V

The classification of movements is made in a hierarchical manner. Indeed a first algorithm classifies the movement into two classes. The first concerns the movements made by the whole body while the second represents the movements made only by hands. In this algorithm we used only five parameters {P2, P3, P4, P5 and P6}. The second algorithm achieves an overall classification and uses the entire set of parameters.

The clustering algorithm K-means (MacQueen, 1967) allows to partition the set of movements into k classes {C1, C2, …, Ck}. U1 partition of the first algorithm contains two rows and n columns. While for the second algorithm U2 contains 6 rows and n columns where n is the number of video sequences. For each sequence a vector V is generated.

$$U1=\begin{bmatrix} u_{1,1} & u_{1,2} & \cdots & u_{1,n} \\ u_{2,1} & u_{2,2} & \cdots & u_{2,n} \end{bmatrix};\ U2=\begin{bmatrix} u_{1,1} & u_{1,2} & \cdots & u_{1,n} \\ u_{2,1} & u_{2,2} & \cdots & u_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ u_{6,1} & u_{6,2} & \cdots & u_{6,n} \end{bmatrix} \tag{10}$$

Where $u_{i,j} \in \{0,1\}$: means the belonging of the movement Pj to the class Ci.

$$\begin{cases} \text{if } P_j \in C_i \ \text{ then } u_{i,j} = 1 \\ \text{else} \quad u_{i,j} = 0 \end{cases} \tag{11}$$

In addition, we impose the following two constraints on each partition

$$\sum_{i=1}^{K} u_{i,j} = 1,\ j = 1,\ldots,N \tag{12}$$

$$\sum_{i=1}^{N} u_{i,j} > 0,\ i = 1,\ldots,K \tag{13}$$

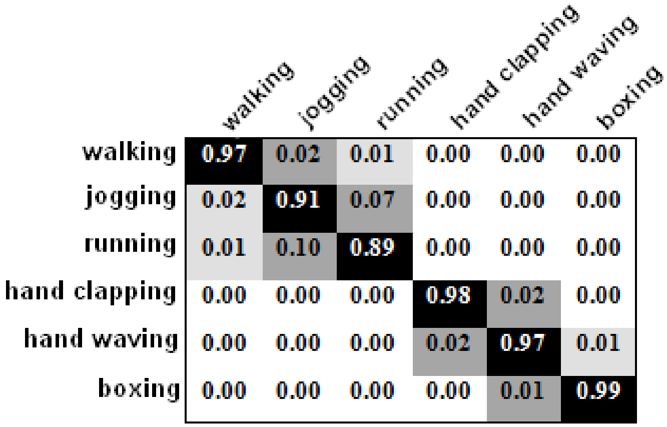
With K is equal to 2 for the first algorithm and 6 for the second. The first specifies that any sample movement must belong to one and only one class of the partition, while the second specifies that a class must have at least one sample of movement.

5. Classification results

The KTH human action database is the largest database available. Each video contains a single action. The database contains six types of human movements (walking, jogging, running, boxing, hand waving and hand clapping). These movements are performed several times by 25 persons in different scenarios, in external or internal environment. The database contains a total of 600 long sequences, that can be divided to more than 10 short sequences of 4 seconds each one.

To test the results of our approach for the recognition task, we used 25% of samples from the video database for the learning task. The 75% remaining video samples are used in the validation task of the performance of the method developed. Figure 10 shows the confusion matrix of classification results for the KTH database.

The confusion matrix in Figure 10 shows the performance obtained for the KTH human action database. Indeed, 450 samples were used to obtain these results (75 for each class). Each column of the matrix represents the accuracy of a class estimated, while each row represents the accuracy of a real class.



	walking	jogging	running	hand clapping	hand waving	boxing
walking	0.97	0.02	0.01	0.00	0.00	0.00
jogging	0.02	0.91	0.07	0.00	0.00	0.00
running	0.01	0.10	0.89	0.00	0.00	0.00
hand clapping	0.00	0.00	0.00	0.98	0.02	0.00
hand waving	0.00	0.00	0.00	0.02	0.97	0.01
boxing	0.00	0.00	0.00	0.00	0.01	0.99

Fig. 10. Confusion matrix for KTH human action database

The best accuracy is obtained for running action while boxing action has the lowest accuracy. The overall recognition rate of our approach exceeds 95%.

The developed approach leads to interesting results compared to other algorithms for human action recognition. All these methods use STIPs to characterize movements without tracking algorithms or background segmentation. Our approach is also comparable to methods based on tracking or segmentation. In Table 4, we illustrate the classification of different approaches according to their accuracy.

Method	Year	Accuracy
Our Method	2011	95,17 %
Xunshi et al.	2010	90,30 %
Ikizler et al.	2009	89,40 %
Niebles et al.	2008	83,33 %
Dollár et al.	2005	81,17 %

Table 4. Classification of different approaches according to their accuracy

6. Conclusion

In this chapter we presented the approach developed for the human action recognition using spatiotemporal interest points STIPs. The STIPs were detected by the application of Laptev STIPs detector. Our classification approach is based on a parameter vector deduced from different studies. The first concerns STIPs number in 100 frames, the second studies the evolution of this number in each frame of the sequence while the third classifies the STIPs in spatiotemporal boxes associated to different parts of the body. For classification we used the k-means classifier. The approach developed has leaded to good performances compared to the well known methods for human action recognition.

As we have only considered K-means as the classification algorithm, we are actually implementing SVM and pLDA algorithms and we plane to make a comparative study between them. Additionally, other metrics will be used to evaluate the methods performances such as Precision, Recall, True Negative Rate (TNR) etc.

7. References

Bobick, A. F. & Davis, J. W. (2001). The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.23, No.3, (Mars 2001), pp. 257–267, ISSN 0162-8828.

Dollár, P.; Rabaud, V.; Cottrell, G. & Belongie, S. (2005). Behavior recognition via sparse spatio-temporal features, *Proceedings of 2nd joint IEEE international workshop on visual surveillance and performance evaluation of tracking and surveillance,* pp. 65–72, ISBN 0-7803-9424-0, Beijing, China, October 15-16, 2005.

Efros, A. A., Berg, A. C., Mori, G., & Malik, J. (2003). Recognizing action at a distance. *Proceedings of the ninth IEEE international conference on computer vision,* Vol. 2, pp. 726–733, ISBN 0-7695-1950-4, Nice, France, October 13-16, 2003.

Harris, C., & Stephens, M. (1988). A combined corner and edge detector. *Proceedings of the fourth Alvey vision conference,* pp. 147– 152, University of Manchester, UK, August 31- September 2, 1988.

Hoey, J. (2001). Hierarchical unsupervised learning of facial expression categories. *Proceedings of IEEE workshop on detection and recognition of events in video,* pp. 99–106, ISBN 0-7695-1293-3, Vancouver, Canada, July 8, 2001.

Ikizler, N., & Duygulu, P., (2009). Histogram of oriented rectangles: A new pose descriptor for human action recognition. *Image and Vision Computing,* Vol.27, No.10, (September 2009), pp. 1515–1526, ISSN 0262-8856.

Kadir, T., & Brady, M., (2003). Scale saliency: a novel approach to salient feature and scale selection. *Proceedings of international Conference on Visual Information Engineering,* pp. 25–28, ISBN 1-55860-715-3, November, 2000.

Koelstra, S., & Patras, I., (2009). The fast-3D spatio-temporal interest region detector. *Proceedings of 10th Workshop on Image Analysis for Multimedia Interactive Services*, pp. 242-245, ISBN 978-1-4244-3609-5, London, UK, May 6-8, 2009.

Laganière, R., Bacco, R., Hocevar, A. Lambert, P. Païs, G. and Ionescu B.E., Video summarization from spatio-temporal features. ACM, 2008.

Laptev, I. (2005). On space-time interest points. *International Journal of Computer Vision*, Vol.64, No.2–3, (September 2005), pp. 107–123, ISSN 0920-5691

Laptev, I., & Lindeberg, T., (2004). Local descriptors for spatiotemporal recognition. Proceedings of First International Workshop "Spatial Coherence for Visual Motion Analysis" Springer LNCS Vol.3667, pp. 91-103, ISBN 3-540-32533-6. Prague, Czech Republic, May, 15, 2004.

Lowe, D., (1999). Object recognition from local scale-invariant features. *Proceedings of International Conference on Computer Vision*, pp. 1150–1157, ISBN 0-7695-0164-8, Kerkyra, Corfu, Greece, September, 20-25, 1999.

MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability,* pp. 281–297, University of California, USA, June 21-July 18, 1965 and December 27, 1965-January 7, 1966

Niebles, J., Wang, H., & Fei-Fei, L., (2008). Unsupervised learning of human action categories using spatial-temporal words, *International Journal of Computer Vision* Vol.79, No.3 (September 2008), pp. 299–318, ISSN 0920-5691.

Oikonomopoulos, A., Patras, I., & Pantic, M., (2006). Spatiotemporal Salient Points for Visual Recognition of Human Actions, *IEEE Trans. Sys. Man. and Cybernetics,* Part B Vol.36, No.3, (June 2006), pp. 710-719, ISSN 1083-4419.

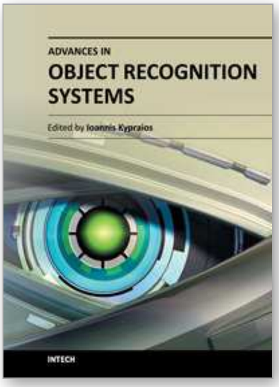
Ramanan, D., & Forsyth, D. A., (2004). Automatic annotation of everyday movements. In: *Advances in neural information processing systems,* Thrun, S.; Saul, L.; & Schölkopf, B., (Eds.), Vol.16, ISBN 0-262-20152-6 Cambridge: MIT Press.

Simac-Lejeune, A., Rombaut, M., & Lambert, P., (2010). Points d'intérêt spatio-temporels pour la détection de mouvements dans les vidéos. *Proceedings of MajecSTIC 2010,* Bordeaux, France, october, 13-15, 2010.

Schuldt, C., Laptev, I., & Caputo, B. (2004). Recognizing human actions: a local svm approach. *Proceedings of the 17th International Conference on Pattern Recognition,* pp. 32–36, ISBN 0-7695-2128-2, Cambridge, England, UK., August, 23-26, 2004.

Xunshi, Y., & Yupin, L. (2010). Making full use of spatial-temporal interest points: an ADABOOST approach for action recognition, *Proceedings of IEEE 17th International Conference on Image Processing*, pp. 4677- 4680, ISBN 978-1-4244-7992-4, Hong Kong, China, September, 26–29, 2010.

Zhong, H., Shi, J., & Visontai,M. (2004). Detecting unusual activity in video. *Proceedings of the 2004 IEEE computer society conference on computer vision and pattern recognition,* pp. 819–826, ISBN 0-7695-2158-4, Washington DC, USA, June 27–July 2, 2004.



Advances in Object Recognition Systems
Edited by Dr. Ioannis Kypraios

ISBN 978-953-51-0598-5
Hard cover, 170 pages
Publisher InTech
Published online 09, May, 2012
Published in print edition May, 2012

An invariant object recognition system needs to be able to recognise the object under any usual a priori defined distortions such as translation, scaling and in-plane and out-of-plane rotation. Ideally, the system should be able to recognise (detect and classify) any complex scene of objects even within background clutter noise. In this book, we present recent advances towards achieving fully-robust object recognition. The relation and importance of object recognition in the cognitive processes of humans and animals is described as well as how human- and animal-like cognitive processes can be used for the design of biologically-inspired object recognition systems. Colour processing is discussed in the development of fully-robust object recognition systems. Examples of two main categories of object recognition systems, the optical correlators and pure artificial neural network architectures, are given. Finally, two examples of object recognition's applications are described in details. With the recent technological advancements object recognition becomes widely popular with existing applications in medicine for the study of human learning and memory, space science and remote sensing for image analysis, mobile computing and augmented reality, semiconductors industry, robotics and autonomous mobile navigation, public safety and urban management solutions and many more others. This book is a "must-read" for everyone with a core or wider interest in this "hot" area of cutting-edge research.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821